Zur Evaluation der Delphi-Technik: eine Ergebnisübersicht

Häder, Michael

Veröffentlichungsversion / Published Version
Arbeitspapier / working paper

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

Empfohlene Zitierung / Suggested Citation:
Häder, M. (1996). *Zur Evaluation der Delphi-Technik: eine Ergebnisübersicht.* (ZUMA-Arbeitsbericht, 1996/02). Mannheim: Zentrum für Umfragen, Methoden und Analysen -ZUMA-. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-200767

ZUMA-Arbeitsbericht 96/02

Zur Evaluation der Delphi-Technik.
Eine Ergebnisübersicht

Michael Häder

April 1996

ZUMA
Quadrat B2,1
Postfach 12 21 55
68072 Mannheim

Telefon: (0621) 12 46 - 136
Telefax: (0621) 12 46 - 100
E-mail: haeder@zuma-mannheim.de

Gliederung

1. Problemstellung: Die Erwartungen an die Delphi-Methode

Die Delphi-Methode verspricht ihrem Anwender, über Sachverhalte, zu denen ihm zunächst nur unsicheres und unvollständiges Wissen vorliegt, zuverlässigere Informationen zu liefern. Dieses Ziel wird durch einen stark strukturierten Gruppenkommunikationsprozeß unter Experten erreicht (vgl. zur Definition von Delphi: Köhler 1992, Linstone/ Turoff 1975, Häder/ Häder 1995 u.a.). Dabei geben die einzelnen Experten zunächst in einer Basiserhebung jeweils ein individuelles Urteil ab, welches in mehreren Wellen über eine anonymisierte Rückmeldung, z.B. in Form eines Mittelwertes der Urteile aller beteiligten Experten, den Teilnehmern bekannt gegeben wird. Nach Kenntnisnahme dieser Gruppenmeinung werden die Experten um ein erneutes Urteil gebeten. Auf diese Weise kommt es zur gezielten Auslösung kognitiver Prozesse (zu den kognitionspsychologischen Grundlagen der Urteilsbildung bei Delphi vgl. z.B. Häder/ Häder 1995) und schließlich zu einer Verbesserung der Qualität der Ausgangsinformation.

Die Delphi-Methode wurde seit ihrer Entwicklung Ende der 40er Jahre bisher zur Aufklärung sehr unterschiedlicher Sachverhalte herangezogen. Diese reichen von den zu erwartenden Ergebnissen eines Hunde- oder Pferderennens (vgl. Woudenberg 1991, S. 132; Seeger 1979, S. 57), über mögliche militärische Angriffsziele (vgl. Linstone/ Turoff 1975, S. 10) bis hin zu langfristigen Voraussagen wissenschaftlicher und technischer Entwicklungen (vgl. z.B. BMFT 1993). Seeger schätzte bereits 1979, es habe innerhalb der 15jährigen Anwendungsdauer ca. 1500 Delphi-Studien verschiedenster Art, insbesondere jedoch in der Betriebswirtschaft, gegeben (vgl. Seeger 1979, S. 32). In den deutschen Sozialwissenschaften wird die Delphi-Methode jedoch bislang kaum angewendet.[1]

Einer der Hauptansatzpunkte von Kritik an der Nutzung von Delphi-Befragungen besteht gegenwärtig in der noch mangelhaften bzw. unvollständigen Evaluation der Qualität der mit diesem Verfahren gewonnenen Ergebnisse. Es liegt die Vermutung nahe, daß ein stärkeres Vertrauen auch der Sozialwissenschaften gegenüber dem Delphi-Ansatz nicht zuletzt durch eine solche Evaluation entwickelt werden kann.

[1] Eine Auflistung von Anwendungsbeispielen der Delphi-Methode findet sich bei Häder/ Häder (1994a).

An dieser Stelle wird zunächst auf einige typische Ansätze bei den bisherigen Evaluationsbemühungen hingewiesen. Danach wird die Methodik solcher Evaluationsansätze kritisch betrachtet. Schließlich sollen die Ergebnisse der Evaluation von zwei konkreten Delphi-Ansätzen vorgestellt werden. Abschließend wird dann nochmals die Frage aufgegriffen, wie eine geeignete Methodik zur Beurteilung von Delphi aussehen kann.

Die Bemühungen um die Überprüfung der Ergebnisse von Delphi-Befragungen lassen sich an folgenden drei typischen Beispielen gut darstellen:

- Zunächst dienten Almanachfragen dazu, um zu prüfen, ob mit Hilfe der Delphi-Methode eine Annäherung an die richtigen Antworten erzielt werden könne. Die RAND Corporation erarbeitete dazu experimentelle Panels (vgl. Dalkey 1969a, Albach 1970, Becker 1974, Geschka 1977). Insgesamt sind solche Versuche jedoch noch nicht sehr zahlreich, sondern bilden eher eine Ausnahme.
- In einem anderen, von der NASA veranstalteten Experiment sollte angegeben werden, welche Ausrüstungsgegenstände die Mannschaft eines Raumschiffes benötigt, das auf dem Mond gestrandet ist (vgl. Erffmeyer/ Lane 1984). Hier zeigte sich, daß mit Hilfe der Delphi- Methode im Vergleich zu anderen Ansätzen die beste Schätzung abgegeben worden ist. Kritik an diesem Experiment äußert allerdings Woudenberg, der darauf verweist, daß es fraglich erscheint, ob das von den NASA-Experten als richtig postulierte Ergebnis auch das tatsächliche ist (1991, S. 136f.).
- Eine andere Evaluationsmöglichkeit bietet inzwischen die Gegenüberstellung von mit der Delphi-Methode gewonnenen Prognose-Aussagen und den tatsächlich eingetretenen Entwicklungen. So wurden beispielsweise 1971 in Japan mögliche Entwicklungslinien von Wissenschaft und Technik vorausgesagt. 20 Jahre danach ließen sich davon bereits 530 Fragen an der Realität beurteilen (zu den insgesamt als positiv eingeschätzten Ergebnissen vgl. im einzelnen: BMFT 1993, S. XXff.).

Kritische Zusammenfassungen verschiedener Versuche, die Delphi-Technik zu evaluieren, finden sich bei Woudenberg (1991) und bei Rowe et al. (1991). Diese Autoren verweisen auf verschiedene generelle Probleme, von denen an dieser Stelle lediglich drei Aspekte aufgegriffen und diskutiert werden sollen.[2]

[2] Weitere von den Autoren genannte Probleme - auf die an dieser Stelle jedoch nur verwiesen werden kann - sind z.B. die bei solchen Tests einbezogene, mitunter zu geringe

Erstens standen zunächst Bemühungen im Mittelpunkt, nach einer a priori Über- bzw. Unterlegenheit der Delphi-Methode gegenüber anderen Verfahren zu suchen. Woudenberg referiert dazu die Ergebnisse von insgesamt 14 Studien. Bei Rowe et al. (1991) werden weitere Untersuchungen ausgewertet. Diese Arbeiten stehen alle vor dem Problem, geeignete Kriterien für eine solche Evaluation zu finden. Woudenberg faßt diese Bemühungen zusammen: „The most feasible way to evaluate the accurracy of Delphi is to compare it directly to other judgment methods in the same situation (1991, S.134)."

Dieser Überlegung folgend stellt er die mit der Delphi-Methode gewonnenen Schätzungen solchen gegenüber, die mit Hilfe anderer Methoden - er betrachtet vor allem unstrukturierte und strukturierte Gruppeninteraktionen - gewonnen worden sind. Ein solches Design erweist sich allein aufgrund des methodischen Variantenreichtums von Delphi-Ansätzen jedoch als ebenfalls nicht unproblematisch, sondern führt zu einem gewissen Dilemma. Dieses bemerkt auch Woudenberg, wenn er resümierend schreibt: „A definite conclusion regarding reliability of the Delphi method must therefore be postponed. The present data, however, do suggest the reliability of *the* Delphi method can hardly be expected to exist" (1991, S. 145, Hervorhebung wie im Original, d. Verf.). Ein ähnliches Ergebnis finden Rowe et al.: „After all, there is no reason why Delphi *should not* be used for aiding forecasts of the near future or assessing present trends for which suitable data may be lacking" (1991, S. 241, - im Original kursiv, d. Verf.).

Ohne das Ergebnis der hier später folgenden Analyse vorwegzunehmen ist festzustellen: Es ist bisher auf diese Weise weder gelungen nachzuweisen, daß die Grundidee des Delphi-Ansatzes anderen Gruppeninteraktionsverfahren generell überlegen ist, noch war es möglich zu zeigen, daß Delphi gegenüber anderen Ansätzen prinzipielle Nachteile aufweist.

Zweitens hat sich als eine Folge dieser Erkenntnis die Diskussion um die Anwendungsberechtigung des Delphi-Verfahrens auf den Kostenaspekt verlagert: „Fischer has pointed out that, when no method is significantly better than any other at providing enhanced judgmental performance, then practical economic factors become important" (Fischer 1981, zitiert nach Rowe 1991, S. 239). Danach spielen - da es nicht mehr um

---

Zahl von Experten, der fast ausschließliche Einsatz von Studenten als Experten, die häufige Nutzung von Laborexperimenten usw.

die grundsätzliche Legitimierung von Delphi geht - nun vor allem solche Aspekte wie der jeweils konkrete Zeit- und Kostenaufwand bei einer Entscheidung für oder gegen den Einsatz des Delphi-Verfahrens eine wesentliche Rolle.

Auch diese Entwicklung kann nicht als befriedigend eingeschätzt werden, da die sich gerade aus dem Variantenreichtum von Delphi ergebenden Möglichkeiten bei einer solchen Betrachtungsweise unberücksichtigt bleiben. Angebracht wäre deshalb die stärkere Berücksichtigung einer Kosten-Qualitäts-Relation in jedem einzelnen Anwendungsfall.

Drittens - und hier knüpft unser Aufsatz an - besteht weiterhin ein beträchtlicher methodischer Forschungsbedarf in der Perfektionierung des jeweils konkret gewählten Delphi-Ansatzes. Da globale (vor allem mit anderen Verfahren vergleichende) Urteile über die Leistungsfähigkeit des Delphi-Ansatzes offenbar fehlgeschlagen sind, geht es nun darum, die mit einem speziellen Delphi-Ansatz gewonnenen Aussagen zu bewerten und nach Möglichkeiten zu suchen, diese zu qualifizieren. Die Attraktivität von Delphi kann dadurch erhöht werden, daß gezeigt wird, auf welche Weise für bestimmte Anwendungsgebiete mit Hilfe bestimmter Modifikationen des Designs valide und zugleich kosten- und zeitgünstige Ergebnisse gewonnen werden können.

2. Evaluation der Delphi-Methode mit Hilfe von Schätzungen der Ergebnisse von Bevölkerungsumfragen

Aufgrund des gewählten Ausgangspunktes, möglichst generelle Aussagen zur Delphi-Methode gewinnen zu wollen, waren - wie im vorigen Abschnitt gezeigt wurde - die bisherigen Evaluationsversuche problembehaftet. Eine eindeutige Antwort auf die selbstgestellte Ausgangsfrage konnte auf diesem Weg nicht gefunden werden: „The only justified conclusion seems to be that factors other than the specific method used (capability of the group leader, motivation of the participants, quality ot the instruction, etc.) to a large extent determine the accuracy of an application of a judgement method“ (Woudenberg 1991, S. 139). Zwei Schlußfolgerungen sind daraus zu ziehen: Erstens sind konkrete Designmodifikationen, die zu einer Verbesserung der Ergebnisse einer Delphi-Befragung führen, zu entwickeln. Zweitens sind anstelle globaler Kriterien zur Beurteilung von Delphi geeignetere Merkmale zu erarbeiten, nach denen solche Designmodifikationen beurteilt werden können. Die Evaluationskriterien verdienen selbst eine Evaluation.

Im weiteren werden Ergebnisse vorgestellt, die bei der Evaluation einer konkreten Anwendung der Delphi-Methode gewonnen wurden. Mit Hilfe der Schätzung von (dem Versuchsleiter bekannten) Antwortverteilungen einer Bevölkerungsumfrage wurde ein spezifisches Anwendungsgebiet der Delphi-Methode in das Zentrum der Betrachtung gestellt. Zugleich besteht damit die Möglichkeit, anhand eines überprüfbaren Kriteriums die Ergebnisse zu evaluieren und Schlußfolgerungen für die Verfeinerung dieser konkreten Nutzungsvariante von Delphi zu ziehen. Ein anderes denkbares Ergebnis wäre natürlich auch eine Empfehlung, für bestimmte Fragestellungen die Delphi-Methode *nicht* einzusetzen. Die folgenden Betrachtungen zur Validierung von Delphi beziehen sich damit ausschließlich auf das hier ausgewählte Anwendungsgebiet, die Schätzung von Antworten von Bevölkerungsbefragungen durch Experten.

In zwei am ZUMA Mannheim bzw. an der Universität Marburg 1994 und 1995 veranstalteten empirischen Tests wurden Experten darum gebeten, mit Hilfe des Delphi-Designs verschiedene Ergebnisse einer Bevölkerungsbefragung zu schätzen. Die abgegebenen Schätzungen wurden den Testteilnehmern in Form von Mittelwerten und Streuungs-

maßen rückgemeldet. Danach ist eine erneute Befragungswelle gestartet worden, in der die Experten ihre Schätzungen wiederholten. Die Erhebungen erfolgten jeweils in drei Wellen. Am ersten Test waren 20 und am zweiten Test 32 Personen beteiligt. In den beiden Tests kamen unterschiedliche Schätzaufgaben zum Einsatz. So sollten jeweils die Randverteilungen von:

(1.) vierstufig skalierten Fragen,

(2.) von Alternativfragen sowie die

(3.) Mittelwerte siebenstufiger Fragen geschätzt werden.

Im zweiten Test war neben diesen Aufgaben außerdem ein Splitt enthalten. Dieser sah in der einen Version vor:

(5.) Antwortverteilungen von fünfstufig skalierten Fragen zu schätzen und in der anderen Version,

(6.) sollte jeweils der Mittelwert dieser beiden Fragen herausgefunden werden.

Beide Tests variierten bei der Erhebung in verschiedenen Details, sodaß ein direkter Vergleich der Ergebnisse nicht möglich ist. Die Resultate der Tests wurden im einzelnen bereits an anderer Stelle ausführlich vorgestellt und dokumentiert (vgl. Häder/ Häder 1994b, Häder/ Häder/ Ziegler 1995).[3] Insgesamt liegen 58 Einzelaufgaben vor, anhand derer die Ergebnisse der Delphi-Befragung bewertet werden können. Für eine solche Bewertung kann jedoch nicht auf verbindliche Kriterien, Validitätskoeffizienten, Standards oder ähnliches, zurückgegriffen werden. Damit müssen zunächst verschiedene, plausibel erscheinende Gesichtspunkte entwickelt und nacheinander besprochen werden.

2.1. Erstes Evaluationkriterium: Fehlerverringerung

Als *erstes Evaluationskriterium* wird die Veränderung der Fehlergröße - hier definiert als Abstand zwischen dem zu schätzenden bzw. dem wahren Wert und dem tatsächlich geschätzten Wert - von Welle zu Welle benutzt. Ein solches Kriterium liegt nahe, da mit Hilfe der Delphi-Methode Gruppenprozesse aktiviert werden sollen, um zunächst naturgemäß unsichere Urteile schrittweise zu qualifizieren. Die Verringerung des Fehlers gibt

[3] Im Anhang werden die in den beiden Delphi-Befragungen geschätzten Indikatoren widergegeben.

damit mehr oder weniger direkt Auskunft über das Gelingen der Expertenbefragung. Noch keine Schlußfolgerung ist anhand dieses Kriteriums jedoch hinsichtlich der Genauigkeit möglich, mit der der wahre Wert schließlich ermittelt worden ist. Eine Übersicht gibt die folgende Tabelle 1.

Tabelle 1: Erfolgsquoten bei der Verringerung der Fehlergröße bei verschiedenen Aufgabentypen innerhalb zweier Delphi-Befragungen

Aufgabentyp	Gesamtzahl der Aufgaben	Erfolgsquote (in Prozent)
4stufige Indikatoren	10	90
2stufige Indikatoren	8	37
Mittelwertschätzungen (7stufig)	36	61
Mittelwertschätzungen (5stufig)	2	100
5stufige Indikatoren	2	50
Gesamt:	58	63

Bei der Betrachtung der Ergebnisse zeigt sich, daß bei 63 Prozent der Aufgaben ein erfolgreicher Verlauf der Schätzungen konstatiert werden kann. Demgegenüber muß von den insgesamt 58 Schätzungen bei 21 Schätzungen eine Vergrößerung des Fehlers festgestellt werden. Diese Erfolgsquote variiert jedoch für die einzelnen Aufgabentypen recht stark, sie liegt - wie aus Tabelle 1 zu entnehmen - zwischen 100 und 37 Prozent. Besonders bei den vierstufigen Indikatoren kam es zu einem erfolgreichen Verlauf der Meinungsbildung bei den Experten, während dagegen die Schätzung zweistufiger Indikatoren am wenigsten gelang.

2.2. Zweites Evaluationskriterium: Treffgenauigkeit

Als *zweites Evaluationskriterium* dient eine Gegenüberstellung der Lage des wahren Wertes und der Streuung der Schätzungen der letzten Welle. Für den Fall, daß der wahre Wert von der Streuung überdeckt wird - konkret soll der Range, also die Spannweite der

Antworten aller Experten betrachtet werden - wird von einem erfolgreichen Verlauf der Delphi-Befragung anhand dieses Kriteriums gesprochen.

Während das erste Kriterium (vgl. Abschnitt 2.1.) auf die (richtige) Richtung der Veränderung der Gruppenmeinung zielte, gibt dieses zweite Kriterium Auskunft über die Genauigkeit, mit der der wahre Wert schließlich getroffen worden ist. In Tabelle 2 sind die Ergebnisse im Überblick dargestellt.

Tabelle 2: In zwei Delphi-Befragungen erzielte Erfolgsquoten, bestimmt aus der Lage des wahren Wertes und der Streuung der Antworten bei verschiedenen Aufgabentypen

Aufgabentyp	Gesamtanzahl der Aufgaben	Erfolgsquote (in Prozent)
4stufige Indikatoren	10	50
2stufige Indikatoren	8	100
Mittelwertschätzungen (7stufig)	36	89
Mittelwertschätzungen (5stufig)	2	0
5stufige Indikatoren	2	100
Gesamt	58	79

Bei der Analyse der Ergebnisse zeigt sich, daß von insgesamt 58 Schätzungen nur in 12 Fällen der Range der Expertenschätzungen nicht den wahren Wert überdeckt. Benutzt man ein stärkeres Kriterium, z.B. den Interdezilbereich, so überdecken immerhin nur in 22 Fällen die Expertenschätzungen nicht den wahren Wert. Dies ergibt bei 79 Prozent (Range) bzw. bei 62 Prozent (bei Betrachtung des Interdezilbereiches) einen erfolgreichen Ausgang der Schätzungen. Auch hier treffen wir wieder auf eine stark unterschiedliche Verteilung der Erfolgsquoten auf die einzelnen Aufgabentypen. Besonders auffällig ist - dies gilt vor allem für die zweistufig skalierten Indikatoren - daß offenbar die zwei Evaluationskriterien voneinander abweichende Einschätzungen der Ergebnisse der beiden Delphi-Befragungen zulassen: Eine relativ hohe Treffgenauigkeit bei einer gleichzeitig nur geringen Verringerung des Schätzfehlers bei der Schätzung der Antworten auf zweistufige Indikatoren sowie eine relativ starke Fehlerminimierung bei einer gleichzeitig

jedoch nur geringen Treffgenauigkeit (vierstufige Indikatoren). Damit liegt ein Vergleich bzw. eine Synthese der beiden Evaluationskriterien nahe.

2.3. Weitere Evaluationskriterien

Mit der Fehlerverringerung und mit der Treffgenauigkeit wurden zunächst zwei zentrale Kriterien für die Beurteilung des Verlaufs von Delphi-Befragungen betrachtet. In diesem Abschnitt sollen die beiden vorliegenden Studien noch auf vier weitere mitunter in der Literatur erwähnte Kriterien zur Beurteilung der Qualität von Delphi-Studien hin untersucht werden:

Erstens wird untersucht, inwieweit das Gruppenergebnis tatsächlich besser ausgefallen ist als das Ergebnis der besten Experten, bzw. wieviele Experten in ihrem individuellen Urteil besser sind als das Gruppenergebnis.

Zweitens wird geprüft, ob sich die Experten bei den wiederholten Schätzungen stärker am zurückgemeldeten Gruppendurchschnitt orientiert haben, oder ob die Annäherung an den wahren Wert stärker war. Es wird also verfolgt, an welcher Marke sich die Schätzergebnisse der jeweils folgenden Welle stärker annähern.

Drittens wird das Delphi-Ergebnis dem Ergebnis einer simulierte Fallstudie gegenübergestellt.

Viertens werden die Konfidenzintervalle um einige Prozentsätze des Originaldatensatzes mit den Ergebnissen der Delphi-Befragungen verglichen.

2.3.1. Vergleich des Gruppenergebnisses mit dem Ergebnis der besten Expertenschätzungen

Die Entwickler des Delphi-Ansatzes gingen davon aus, daß in n+1 Kopf mehr Wissen vorhanden wäre als in einem. Demzufolge sei zu erwarten, daß das arithmetrische Mittel des Gruppenergebnisses mindestens genauso gut ausfallen werde - wahrscheinlich jedoch besser - als die Einzelergebnisse. Diese Überlegung fortsetzend wird nun diskutiert, inwieweit das Gruppenergebnis auch besser ausfällt als die Schätzergebnisse der besten

Experten. Mit Hilfe eines solchen Kriteriums läßt sich die Effektivität der *Gruppen*leistung darstellen. Es ist damit, anders als die bisher behandelten Aspekte, kein direkter Ausdruck für die Qualität der Schätzergebnisse. Ideal wäre danach die Effektivität einer Delphi-Studie, wenn kein Experte in seiner Schätzung besser läge als das Resultat der gesamten Gruppe. Ein solcher Fall würde bedeuten, daß das positive Ergebnis (lediglich) aufgrund der Gruppenleistung zustandegekommen ist. Dagegen würde es den gewählten Ansatz infragestellen, wenn über 50 Prozent der Experten besser zu schätzen vermögen als die gesamte Gruppe. Läge diese Konstellation vor, wäre die Konsultation einer Gruppe überflüssig - da ohnehin die Mehrheit der Experten bei ihren Schätzungen zu einem richtigen Ergebnis käme.

In Tabelle 3 sind die Ergebnisse der beiden bisher diskutierten Tests enthalten.

Tabelle 3: Prozentsatz der Experten, deren individuelle Urteile *besser* waren als das statistische Gruppenurteil in der dritten Welle, dargestellt für die jeweiligen Aufgabentypen, Grundlage sind zwei separate Befragungen

Aufgabentyp	Anzahl der Aufgaben	Anzahl der Experten	bessere Urteile bei:
4stufige Indikatoren	10	52	33%
2stufige Indikatoren	8	52	23%
Mittelwertschätzungen (7stufig)	36	52	41%
Mittelwertschätzungen (5stufig)	2	16	31%
5stufige Indikatoren	2	16	41%
Gesamt	58	188	31%

Die Ergebnisse sind wiederum bei den einzelnen Aufgabentypen stark unterschiedlich ausgefallen. Am effektivsten war die Gruppenleistung im Vergleich zu den Einzelleistungen offenbar bei den Aufgaben, die die Schätzung der Ergebnisse zweistufiger Indikatoren beinhaltete, hier lag auch die Treffgenauigkeit besonders hoch. Den geringsten Effekt hatte die Gruppenleistung dagegen bei den Mittelwertschätzungen. Dies gilt auch für die offenbar besonders anspruchsvollen Schätzungen der Ergebnisse von fünfstufigen Indikatoren.

Insgesamt liefert diese Übersicht einen weiteren konkreten Hinweis zur Diskussion um die Leistungsfähigkei der Delphi-Methode. Rowe et al. stellen unter Verweis auf Hill (1982) und Hastie (1986) fest:

„Further, studies have generally shown group judgement to be mainly inferior to the judgement of the group's best member" (1991, S. 236).

Vor dem Hintergrund der hier gezeigten Daten verdient diese globale Feststellung jedoch, präzisiert zu werden. So war zwar - dies geht aus der Tabelle 3 nicht im einzelnen hervor - nur in relativ wenigen Fällen *kein* Experte besser als das Gruppenergebnis, jedoch fällt das Gruppenergebnis in der Regel sogar deutlich besser aus als die Mehrheit der Einzelergebnisse. Eine weitere Schlußfolgerung, die aus diesem Ergebnis gezogen werden kann, ist jedoch auch, daß die Gruppenleistung noch nicht das optimale Niveau erreicht hat und in vielen Fällen noch weitere Verbesserungen der Qualität der Schätzurteile zumindest theoretisch denkbar sind - ausgelöst durch ein weiterentwickeltes methodisches Design.[4]

2.3.2. Bewegung zum Gruppendurchschnitt oder zum wahren Wert

In der Literatur werden verschiedene weitere Erwartungen mit dem Einsatz der Delphi-Methode verknüpft. So existiert unter anderem die Hoffnung, daß sich die an einer Delphi-Befragung beteiligten Experten bei der Wiederholung ihrer Urteile stärker am wahren Wert orientieren als an den ihnen rückgemeldeten Gruppenergebnissen aus der Vorrunde. Dabei wird zunächst jedoch zumeist stillschweigend unterstellt, daß letztlich eine mehr oder weniger große Differenz zwischen dem rückgemeldeten Gruppendurchschnitt und dem zu schätzenden wahren Wert besteht. Nur wenn - ein solcher Fall einmal unterstellt - bei der erneuten Urteilsbildung dann eine stärkere Orientierung am wahren Wert erfolgt, kann der Einsatz der Delphi-Methode als Erfolg bewertet werden. Letztlich wird bei einer solchen Bewertung des Delphi-Verfahrens der Konformitätsdruck einerseits der Orientierung am wahren Wert andererseits gegenübergestellt.

[4] Rowe et al. schlagen deshalb beispielweise vor, daß die Experten auch verbale Begründungen für ihre Urteile an die anderen Teilnehmer rückmelden, um so weitere kognitive Prozesse auszulösen. Rowe et al.: „By limiting the scope of feedback one limits the scope for improvements" (1991, S. 244).

Diesem Kriterium wird im weiteren nachgegangen. Tabelle 4 enthält dazu eine entsprechende Zusammenstellung der Ergebnisse.

Tabelle 4: Annäherung an den wahren Wert und Annäherung an den nach der ersten bzw. zweiten Welle rückgemeldeten Gruppendurchschnitt, Ergebnisse von zwei Delphi-Befragungen, dargestellt nach Aufgabentypen

Aufgabentyp	Anzahl der Aufgaben	stärkere Orientierung am: - Gruppendurchschnitt	- wahren Wert
4stufige Indikatoren	80	70	10
2stufige Indikatoren	16	14	2
Mittelwertschätzungen (7stufig)	72	58	14
Mittelwertschätzungen (5stufig)	4	3	1
5stufige Indikatoren	20	14	6
Gesamt	192	149	33

In beiden hier besprochenen Experimenten läßt sich zunächst tatsächlich deutlich eine stärkere Bewegung der Antworten in Richtung auf den Gruppendurchschnitt der vorherigen Welle als in Richtung auf den wahren Wert erkennen. Die von Welle zu Welle eintretende Konvergenz der Antworten erfolgt nicht primär in Richtung auf den wahren Wert, jedoch - diese Interpretation der Ergenisse erscheint berechtigt - auch nicht zwangsläufig in jedem Fall auf den rückgemeldeten Gruppendurchschnitt. Wichtig erscheint allerdings der Hinweis, daß dieses Ergebnis hier völlig unabhängig von der Größe der Differenz zwischen wahrem Wert und Gruppendurchschnitt ermittelt wurde.

Eine Legitimation des Delphi-Ansatzes bei Betrachtung lediglich dieses Evaluationskriteriums könnte aufgrund der beiden hier betrachteten Tests nicht abgeleitet werden. Damit steht dieses Kriterium zumindest in einem formalen Widerspruch zu den bisher betrachteten. Auf dieses Ergebnis wird in den abschließenden Betrachtungen (vgl. Abschnitt 3.1.) noch gesondert eingegangen.

2.3.3. Die Delphi-Methode in einer Gegenüberstellung zu einer simulierten Fallstudie

Es wurde bereits einleitend darauf verwiesen, daß es bisher für die Bewertung der Ergebnisse von Delphi-Studien keine feststehenden, allgemein akzeptierten Kriterien gibt. Dementsprechend wurden in den vorangegangenen Abschnitten einzelne, besonders plausibel erscheinende Gesichtspunkte herausgegriffen und danach die Ergebnisse unserer beiden Tests beurteilt (vgl. die Abschnitte 2.1., 2.2., 2.3.1. und 2.3.2.). Ein anderes zur Evaluation von Delphi-Befragungen bisher mitunter benutztes Vorgehen besteht darin, die mit der Delphi-Methode gewonnenen Ergebnisse mit denen eines anderen Verfahrens zu vergleichen. Für solche Gegenüberstellungen wurden bisher insbesondere Gruppeninteraktionsverfahren herangezogen. (Auf die dabei auftretenden methodischen Probleme wurde im Abschnitt 2. verwiesen.) Zur Illustration der Grenzen eines auf einen Vergleich ausgerichteten Designs und gleichzeitig zur Vervollständigung unserer Darstellung von Bewertungsansätzen soll nun ebenfalls der Versuch unternommen werden, die Ergebnisse der Delphi-Befragung mit denen aus einem anderen Ansatz zu konfrontieren. Aus der Vielzahl der dafür infragekommenden Methoden ist an dieser Stelle exemplarisch die Fallstudie ausgewählt worden. Dabei wird im weiteren die Frage untersucht, inwieweit die mit Hilfe des Delphi-Ansatzes aus der wiederholten Befragung von 16 Experten gewonnenen konkreten Ergebnisse denen ähnlich sind, die aus der Befragung einer etwa gleich großen Anzahl von Vertretern der Grundgesamtheit gewonnen wurden.

Folgendes Vorgehen ist dafür gewählt worden: Anhand ausgewählter Kriterien wurden dem Umfang unseres Expertengreminus entsprechend kleine Subpopulationen aus der ursprünglichen Stichprobe extrahiert. Die Ergebnisse der Befragung dieser Untergruppen wurden dann mit denen in der Delphi-Befragung gewonnenen verglichen.[5]

[5] Dazu waren einigen Konventionen erforderlich. Erstens: Die Bildung der Subpopulation kam durch eine relativ willkürliche Homogenisierung des Originaldatensatzes der Shell-Jugendstudie (vgl. Jugend ‘92) zustande. Als Kriterien dienten das Bundesland und die Tätigkeit der Zielperson. So sind westdeutsche (einschließlich West-Berliner) Studenten aus einem Bundesland und aus möglichst dem gleichen Semester gezielt ausgesucht worden. Dies entspricht zugleich in etwa auch der Struktur und der Größe der bei unserem Delphi-Test befragten Personengruppe.

Zweitens mußten weiterhin die zu betrachtenden Indikatoren festgelegt werden. Um den Aussagewert des Beispiels zu erhöhen sind gezielt jene ausgewählt worden,

Tabelle 5 enthält die Ergebnisse dieses Tests.

Tabelle 5: Vergleich der bei einer repräsentativen Auswahl westdeutscher Jugendlicher zwischen 18 und 29 Jahren (Shell-Studie), bei zwei ausgesuchten Subpopulationen dieser Studie und bei einer Delphi-Befragung gewonnenen Ergebnisse zu zwei Indikatoren

Indikator (1): „Wie stehst Du zur Vereinigung von ehemaliger DDR und alter Bundesrepublik von heute aus gesehen?“

	Shell-Studie	Subpop. 1	Subpop. 2	Delphi
1 sehr dafür	21%	0	32%	11%
2 eher dafür als dagegen	32%	35%	37%	28%
3 unentschieden	24%	12%	16%	32%
4 eher dagegen als dafür	18%	41%	10%	20%
5 sehr dagegen	5%	12%	5%	9%
Mittelwert	2,5	3,3	2,2	2,8
n:	1914	17	19	16

Indikator (2): „Die Menschen sind ja sehr unterschiedlich, wenn es um ihre Lebensziele geht. Manche sind sehr anspruchsvoll und ehrgeizig, andere finden diese weniger gut oder wichtig. Wie ist das bei Dir?
Ich bin bei den Zielen, die ich mir für mein Leben setze ...“

	Shell-Studie	Subpop. 1	Subpop. 2	Delphi
1. nicht so anspruchsvoll oder ehrgeizig	1%	0	0	11%
2.	8%	0	5%	21%
3.	32%	18%	16%	32%
4.	42%	65%	37%	23%
5. sehr anspruchsvoll und ehrgeizig	17%	17%	42%	13%
Mittelwert	3,7	4,0	4,2	3,4
n:	1914	17	19	16

Subpop. 1: West-Berliner Studenten
Subpop. 2: Studenten aus Bayern im zweiten Semester

deren Schätzung im Delphi-Experiment nicht besonders erfolgreich verlaufen ist (vgl. Häder/ Häder/ Ziegler 1995).

Dabei zeigen sich folgende Resultate:

- Beim Delphi-Ansatz gelingt bei beiden Indikatoren eine deutlich bessere Schätzung der fünfstufigen Antwortverteilungen. Insbesondere die Ergebnisse der zweiten Subpopulation weichen beim zweiten Indikator deutlich vom Gesamtergebnis der Shell-Studie ab. So beträgt die Differenz zwischen dem Ergebnis der Delphi-Befragung und dem wahren Wert 0.3 und die Differenz zwischen dem Ergebnis der Subpopulation 2 und dem wahren Wert 0.5.
- Gleichwertige bis ebenfalls leicht bessere Ergebnisse können bei der Schätzung der beiden Mittelwerte im Delphi-Ansatz festgestellt werden (vgl. Tabelle 6).

Tabelle 6: Differenzen zwischen dem wahren Wert und den Ergebnissen der Delphi-Erhebung sowie der beiden gebildeten Subpopulationen der Shell-Studie, Mittelwertvergleich

	Shell - Delphi	Shell - Subpop. 1	Shell - Subpop. 2
Indikator 1	0.3	0.8	0.3
Indikator 2	0.3	0.3	0.5

Bei der Bewertung dieses Ergbnisses ist außerdem zu berücksichtigen, daß bei der Befragung kompetenterer Experten (aus bestimmten Gründen wurden in unserem Test mit empirischer Sozialforschung nur wenig vertraute Medizinstudenten eines ersten Studienjahres befragt) sowie bei einer weniger komplizierten Aufgabenstellung der Vergleich zwischen Delphi und Fallstudie noch stärker zugunsten des Delphi-Ansatzes ausfiele.

Insgesamt besitzt ein solches Vorgehen vor allem eine illustrative Funktion. Aufgrund des Variantenreichtums sowohl von Delphi-Befragungen als auch von Fallstudien ermöglicht die hier gezeigte Gegenüberstellung lediglich einen Einblick in diese Art der Beurteilung des Delphi-Ansatzes. Für detailliertere Aussagen über die Möglichkeiten der Delphi-Methode im Vergleich zu denen von Fallstudien sind indes weitere Untersuchungen erforderlich:

- So könnte in einem Splitt der eine Teil einer Expertengruppe tatsächlich nach deren persönlichen Meinungen zu den Indikatoren befragt werden, während der andere Teil dieser Gruppe die Antworten mit Hilfe eines Delphi-Designs schätzt.

- In die Betrachtung müßten weitere, auch unterschiedliche, Schätzaufgaben einbezogen werden. Außerdem kann der Aussagewert eines solchen Vergleichs durch die Beteiligung einer größeren Anzahl an Vergleichsgruppen weiter erhöht werden.

2.3.4. Vergleich der Konfidenzintervalle mit den Ergebnissen der Delphi-Methode

Werden aus den Ergebnisse einer Stichprobe Prozentsätze hochgerechnet, so sind diese Schätzer für die Grundgesamtheit von einem Konfidenzintervall umgeben. Die Größe dieses Intervalls läßt sich bestimmen aus dem Stichprobenumfang und einer vorgegebenen Irrtumswahrscheinlichkeit (zumeist 5 oder 1 Prozent). Für die Beurteilung der Ergebnisse einer Delphi-Befragung soll nun untersucht werden, in welcher Beziehung das Vertrauensintervall der Originaldaten und die von den Experten geschätzen Ergebnisse stehen. Tabelle 7 zeigt eine entsprechende Gegenüberstellung dieser Ergebnisse für die (Alternativ-)Fragen 8 bis 11.

Tabelle 7: Darstellung der Konfidenzintervalle einiger Ergebnisse der Shell-Studie sowie der in zwei Delphi-Befragungen geschätzten Ergebnisse

Frage	Prozent-wert (Shell)	n	Konfidenzintervalle 95%	99%	Delphi-Schätzung
8 (A)	45	2000	42.8 ... 47.2	42.1 ... 47.9	42 (n=32)
	50	4000	48.5 ... 51.6	48.0 ... 52.0	41 (n=20)
9 (A)	41	2000	38.8 ... 43.2	38.2 ... 43.8	55 (n=32)
	37	4000	35.5 ... 38.5	35.0 ... 39.0	48 (n=20)
10 (A)	63	2000	61.9 ... 65.1	61.2 ... 65.8	65 (n=32)
	62	2000	60.5 ... 63.5	60.0 ... 64.0	64 (n=20)
11 (A)	45	2000	42.8 ... 47.2	42.1 ... 47.9	55 (n=32)
	43	4000	41.5 ... 44.5	41.0 ... 45.0	43 (n=20)

Aufgrund des relativ großen Stichprobenumfangs sind die Konfidenzintervalle relativ klein. Das Ergebnis in Tabelle 7 zeigt, daß trotzdem etwa die Hälfte der Delphi-Schätzungen im Konfidenzintervall der Originaldaten liegt. Die von den Experten zu schätzenden (wahren) Werte sind selbst mit einem gewissen Fahler behaftet. Das in einer

Delphi-Befragung gewonnene Ergebnis kann deshalb auch mit dem Vertrauensintervall des wahren Wertes in Beziehung gesetzt werden. Damit steht dann ein weiteres Kriterium für die Evaluation des gesamten Schätzerfolges zur Verfügung.

3. Beurteilung der Evaluationskriterien

Die Ergebnisse erlauben es, auf die zwei bereits im Abschnitt 1. genannten Aspekte der Evaluation des Delphi-Ansatzes nochmals zurückzukommen:

Erstens wird dem Delphi-Verfahren von Kritikern unterstellt, daß es lediglich Ergebnisse produziere, die sich, ausgehend von Konformitätsdruck, auf eine Annäherung an das Gruppenmittel zurückführen lassen und damit nicht zwangläufig zu einer Lösung des Problems führen müssen. So schätzt beispielsweise Brockhoff im Schlußwort seiner Darstellungen ein: „Die Veränderungen der Schätzungen im Verlauf der Delphi-Prozesse läßt zwar Konvergenz der Urteile erkennen, doch scheint die Bewegung zum Median der Vorrunde im allgemeinen stärker als die Bewegung der Urteile zum wahren Wert“ (1979, S. 165, eine ähnliche Auffassung vertreten auch Seeger (1979, S. 212) und Woudenberg (1991, S. 145)). Auch in beiden hier betrachteten Tests - vgl. Abschnitt 2.3.3. - dominierte die Annäherung an den rückgemeldeten Wert. Welche Aussagekraft besitzt ein solches Kriterium jedoch, wenn man es im Zusammenhang mit den anderen hier behandelten betrachtet?

Zweitens wird die Leistungsfähigkeit der Delphi-Methode mit der anderer Verfahren (z.B. face-to-face Gruppendiskussionen) verglichen und daraus mitunter die Schlußfolgerung abgeleitet, daß ein Verfahren das andere ersetzen solle, bzw. zunächst die Überlegenheit von Delphi gegenüber der Gruppendiskussion nachgewiesen werden müsse (vgl. Seeger 1979, S. 150), bevor eine Anwendung legitimiert sei. Läßt sich eine solche Haltung auch vor dem Hintergrund der referierten Ergebnisse (z.B. zu Fallstudien) weiter vertreten?

3.1. Einschätzung der Gruppenleistung

Die in den vorigen Abschnitten referierten Ergebnisse zur Evaluation der gewählten speziellen Anwendungsform einer Delphi-Befragung sind offenbar inkonsistent und verdienen es damit, ausführlicher diskutiert zu werden. So ist beispielsweise die Schätzung der Antworten von *vierstufigen Skalen* anhand des Kriteriums „Verringerung der Fehler-

größe“ besonders erfolgreich verlaufen (vgl. Tabelle 1). Wählt man dagegen das Kriterium „Treffgenauigkeit“ (vgl. Tabelle 2), so kommt man zu dem Schluß, daß hier das Delphi-Verfahren nicht besonders überzeugend funktioniert hat. Wieder ein anderes Bild zeigt Tabelle 3, in der die Gruppenleistung mit der Leistung der besten Experten verglichen wird. Hier konnte bei diesem Aufgabentyp wiederum ein recht gutes Ergebnis festgestellt werden. Schließlich fällt eine Gegenüberstellung der Orientierung der Experten am wahren Wert bzw. am rückgemeldeten Mittelwert auch bei diesem Aufgabentyp relativ unbefriedigend aus (vgl. Tabelle 4).

Eine These, nach der sich die Ergebnisse von Delphi-Befragungen *primär* aufgrund einer Tendenz zur Annäherung an den Gruppendurchschnitt erklären lassen, bzw. derzufolge dieses Kriterium eventuell sogar zu einem Schlüsselgesichtspunkt für die Beurteilung von Delphi-Befragungen erklärt wird, kann durch unsere Ergebnisse empirisch nicht unterstützt werden. Plausibel und mit den vorgestellten Daten vereinbar erscheinen dagegen die beiden folgenden kognitionspsychologischen Argumentationsketten:

1. Eine mögliche Ausgangsannahme besteht zunächst darin, daß die Experten dazu in der Lage sind, die Zuverlässigkeit des eigenen Urteils mehr oder weniger richtig abzuschätzen. So wurde in einem Experiment festgestellt, daß Teilnehmer, die sich selbst stärker für kompetent einstufen, tatsächlich auch bessere Schätzungen abgeben als Teilnehmer, die für sich einen geringeren Grad an Experteneigenschaft angaben (vgl. dazu: Dalkey 1969a, Geschka 1977, Albach 1970, Becker 1974). Experten verfügen also über ein gewisses Maß an Wissen darüber, mit welcher Wahrscheinlichkeit die von ihnen benutzten Erkenntnishilfen (in der Kognitionspsychologie als „cue-Validität“ bezeichnet[6]) auf den gesuchten Sachverhalt verweisen.[7] Daraus ergeben sich zwei weitere Vermutungen: *Erstens* wird die erhaltene Rückinformation in Abhängigkeit von diesem internen Qualitätsmaßstab bewertet. So wird die Gruppenmeinung wahr-

[6] Der Begriff Cue-Validität meint die „subjektive Repräsentation des Zusammenhangs zwischen Cue und Zielvariable; der Begriff der ökologischen Validität hingegen verweist auf den tatsächlichen Zusammenhang.“ (Hoffrage 1993, 79)

[7] Mitunter wird in Delphi-Befragungen diese von den Experten vermutete eigene Kompetenz mit erhoben. So wurden die im Rahmen eines Delphi-Ansatzes befragten Experten dazu aufgerufen, ihre eigene Kompetenz als „groß“, „mittel“, „gering“ oder „fachfremd“ einzustufen. Letztere brauchten die entsprechenden Fragen nicht zu beantworten (vgl. Cuhls et al. 1995, S. 8f.).

scheinlich anders zur Kenntnis genommen und entsprechend verarbeitet, wenn das eigene Urteil als noch unsicher angesehen wird, als wenn dieses als bereits relativ endgültig bzw. fertig eingeschätzt wird. *Zweitens* bestimmt der Abstand der eigenen Ansicht zur Gruppenmeinung auch den Grad an Bereitschaft, das eigene Urteil zu revidieren bzw. von den anderen Experten zu lernen. Diese Bereitschaft wird dann relativ groß sein, wenn das erste Urteil zunächst auf minderwertigeren cues aufgebaut wurde. Im Falle einer (von den Experten wahrscheinlich zurecht vermuteten) zuverlässigen Ausgangsschätzung ist diese Bereitschaft jedoch relativ gering. Es erfolgt in diesem Fall dann kaum eine primäre Orientierung an der Gruppenmeinung, wie von einigen Kritikern behauptet wird, vielmehr wird diese eher ignoriert und das ursprüngliche Urteil erneut in gleicher Form abgegeben.

2. Eine weitere Möglichkeit für die Erklärung der bei einer Delphi-Befragung ablaufenden Prozesse bieten kognitionspsychologische Überlegungen und Experimente zum Frequency-Validity-Effekt[8]. Für die Erklärung von kognitiven Prozessen, die zu einer identischen Reproduktion eines zunächst abgegebenen Konfidenz-Urteils führen schreibt Hertwig (1993, S. 55 - kursiv wie im Original, d. Verf.): „Mißlingt also die direkte Erinnerung, dann bemüht sich die Versuchsperson, die Wissensbasis zu errichten, die dem ersten Konfidenz-Urteil zugrunde lag, d.h., sie *konstruiert erneut ein mentales Modell.* Gelang es ihr bereits in der ersten Sitzung, ein LMM (Lokales Mentales Modell, d. Verf.) zu bilden, ist es wahrscheinlich, daß das notwendige Wissen oder die notwendigen Deduktionen auch jetzt wieder verfügbar sind, so daß das zweite Konfidenz- und das Erinnerungs-Urteil mit dem ersten Konfidenz-Urteil identisch sind. Gelang hingegen lediglich die Bildung eines PMM (Probabilistisches Mentales Modell, d. Verf.), dann ist es nicht sicher, daß exakt das Erst-Urteil reproduziert werden kann.“

Diese Feststellung beinhaltet eine auf die Delphi-Methode übertragbare wesentliche Aussage - obwohl es sich hier nicht um Konfidenz-Urteile handelt. Sie enthält letztlich eine Begründung dafür, daß bei der Wiederholung einer Schätzung sichere Urteile identisch reproduziert werden, während Urteile, die (lediglich) auf probabilistischen mentalen Modellen beruhen, erneut generiert werden müssen. Letztere haben damit die

[8] Der Frequency-Validity-Effekt bezeichnet die Erscheinung, daß bei der Wiederholung einer Aussage diese von den Testteilnehmern für wahrscheinlich wahrer gehalten wird.

Chance, besser auszufallen als die ursprüngliche Schätzung - wie empirisch in unseren Experimenten dann auch tatsächlich nachgewiesen wurde - z.B. aufgrund der Nutzung zuverlässigerer cues, oder solcher, die bisher nicht berücksichtigt worden sind.

Hertwig schreibt dementsprechend auch weiter: (1993, S. 56; kursiv wie im Original, d. Verf.): „Unterschiede zwischen dem Urteil der ersten und zweiten Sitzung können durch zwei Prozesse - *Berücksichtigung neuer Information* oder *erneute Konstruktion eines mentalen Modells* - erklärt werden.“[10]

Es ist nicht hinreichend, lediglich die Annäherung an den wahren Wert in das Verhältnis zur Annäherung an den rückgemeldeten Gruppendurchschnitt zu setzen, um Delphi-Befragungen zu beurteilen. Als ein „globales Leistungskriterium“ taugt dieser Aspekt nicht, da selbst bei einer Annäherung (nur) an den Gruppendurchschnitt eine Verringerung des Schätzfehlers erfolgen kann. Die Aussagekraft dieses Beurteilungskriteriums ist auch dann besonders eingeschränkt, wenn der Abstand zwischen der zusammengefaßten Gruppenmeinung und dem wahren Wert bereits in einer der ersten Schätzungen sehr gering ist.

Das Problem soll am folgenden Beispiel weiter verdeutlicht werden. Zunächst sind prinzipiell drei Konstellationen zwischen wahrem Wert, dem Gruppendurchschnitt und einer individuellen Schätzung denkbar:

- erstens, eine individuelle Schätzung liegt zwischen dem wahren Wert und dem Gruppenurteil;
- zweitens, die individuelle Schätzung liegt jenseits des Gruppenurteils und
- drittens, die individuelle Schätzung liegt jenseits des wahren Wertes (vgl. Abbildung 1).

---

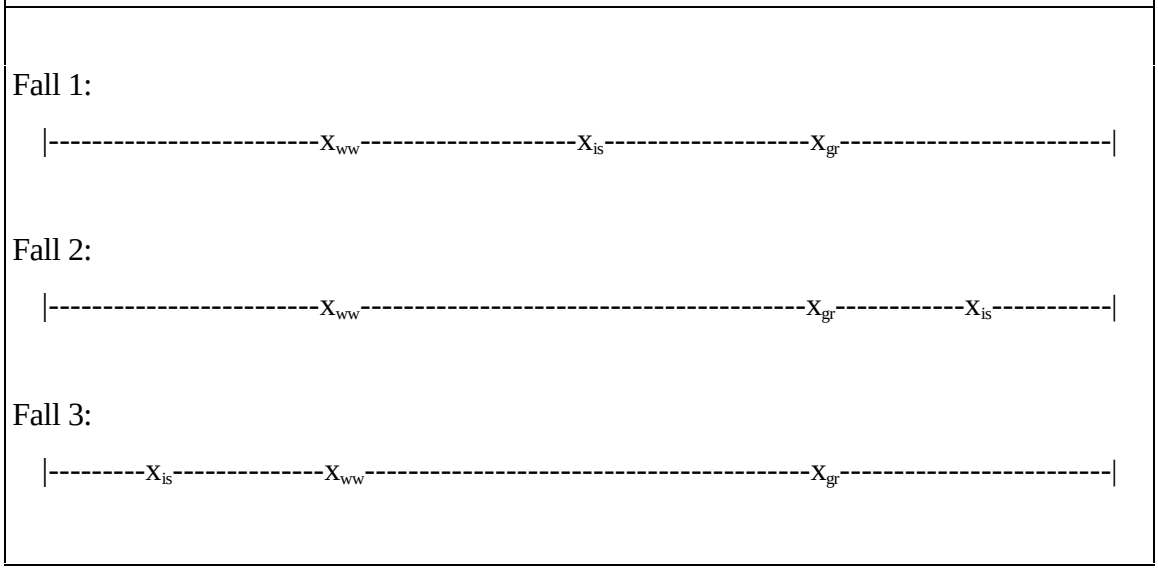
[10] Hertwig setzt seine Argumentation schließlich folgendermaßen fort:

„Bedient man sich ... nur des Cues mit der höchsten Validität, werden die neuen Cues nur dann relevant, wenn sich unter diesen einer befindet, der den bisherigen ‘Spitzenreiter’ überflügeln kann. Liegt hingegen ein Mittlungs-Algorithmus (= mehrere Cues werden in einem Urteil verarbeitet) zugrunde, wird jeder neu generierte Cue (ungeachtet seiner Cue-Validität) einen Einfluß auf das zweite Urteil besitzen“ (1993, 56).

Wichtig erscheint es hier zu betonen, daß es aber auch (z.B. bei einer Delphi-Befragung aufgrund der Rückmeldung) „zu einer Neueinschätzung der Cue-Validität und der Anwendbarkeit des Cues kommen“ kann.

Im ersten Fall bedeutet eine Annäherung an den Gruppendurchschnitt immer auch eine Verschlechterung der Schätzung. Im zweiten Fall bedeutet eine Annäherung an den Gruppendurchschnitt bereits zwangsläufig eine Verbesserung der Schätzung, wobei prinzipiell auch eine noch stärkere Annäherung an den wahren Wert und damit eine weitere Verbesserung des Urteils denkbar wäre. Auch im dritten Fall kann eine Annäherung an den Gruppendurchschnitt eine Verbesserung der Schätzung bedeuten. Lediglich bei einer extremen Orientierung an der Gruppenmeinung würde es jedoch zu einer Verschlechterung der Schätzung kommen.

Abbildung 1: Mögliche Konstellationen zwischen wahrem Wert (x_{ww}), dem Gruppendurchschnitt einer Schätzung (x_{gr}) sowie einer individuellen Schätzung (x_{is})



Aussagekräftiger als die Auswahl lediglich eines Aspekts verspricht dagegen eine Kombination von Kriterien zu sein, um den Erfolg von Delphi-Erhebungen zu beurteilen.

3.2. Erfolgstypen: Kombination von Fehlerverringerung und Treffgenauigkeit

Die beiden Evaluationskriterien Fehlerverringerung (vgl. Abschnitt 2.1.) und Treffgenauigkeit (vgl. Abschnitt 2.2.) zusammenfassend, läßt sich die folgende in Abbildung 2 dargestellte Typologie von möglichen Verläufen bei einer Delphi-Befragung aufstellen.

Danach ergibt sich aus der Kombination der jeweils alternativen Merkmale eine Vier-Felder-Tafel mit folgenden Erfolgs-Typen.

Abbildung 3: Theoretisch gebildete Erfolgstypen von Delphi-Befragungen anhand der Kombination der Evaluationskriterien „Fehlerverringerung“ und „Treffgenauigkeit“

Fehlerverringerung ...	Treffgenauigkeit: - innerhalb des Ranges	- außerhalb des Ranges
- ja	**Erfolg** (Typ 1)	**spezifischer Delphi-Erfolg** (Typ 2)
- nein	**unspezifischer Delphi-Erfolg** (Typ 3)	**Mißerfolg** (Typ 4)

Von „Erfolg“ (Typ 1) wird gesprochen, wenn sich nach der letzten Welle sowohl der ursprünglich in der ersten Welle aufgetretene Fehler verringert hat als auch die Streuung der Urteile in der letzten Welle den wahren Wert überdeckt. Entsprechend soll von „Mißerfolg“ (Typ 4) gesprochen werden, wenn sich sowohl der Fehler in der dritten Welle gegenüber der ersten Welle vergrößert hat und zugleich der wahre Wert außerhalb der Spannweite der Expertenschätzungen der letzten Welle liegt. Genau in diesem Fall hat der Einsatz der Delphi-Methode zu keinem verwertbaren Ergebnis geführt.

Ein (nur) „spezifischer Delphi-Erfolg“ (Typ 2) liegt dann vor, wenn sich zwar der Fehler in der dritten Welle gegenüber der ersten verringert hat, aber die Expertenmeinungen nicht den wahren Wert überdecken. Dies mag beispielsweise dann der Fall sein, wenn es sich um eine besonders schwierige Aufgabe handelt, die nun zwar besser als vor Beginn der Delphi-Befragung gelöst werden kann, die Schätzung jedoch immer noch mit einem relativ hohen Maß an Ungenauigkeit verbunden ist. Eine solche Konstellation stellt ebenfalls einen spezifischen Erfolg dar, da durch die Delphi-Befragung ein letztlich doch erfolgreicher Annäherungsprozeß an den wahren Wert ausgelöst worden ist.

Von einem (lediglich) „unspezifischen Delphi-Erfolg“ (Typ 3) soll dann gesprochen werden, wenn sich die Fehlergröße in der letzten Welle gegenüber der ersten zwar

vergrößert hat, der wahre Wert jedoch von den Expertenurteilen überdeckt wird. In diesem Fall hat die Konsultation der Expertengruppe zwar ebenfalls zu einem Informationsgewinn geführt, jedoch wäre die - für die Delphi-Technik typische - Wiederholung der Befragung nicht erforderlich gewesen. Damit scheint die hier benutzte operationale Bezeichnung „unspezifischer Delphi-Erfolg“ angebracht, da der eingetretene Gewinn an Information nicht auf die Delphi-Technik, sondern auf die Konsultation einer Expertengruppe zurückgeführt werden kann.

Die beiden zuletzt besprochenen Typen (2 und 3) unterscheiden sich von Typ 4 dadurch, daß bei jedem einzelnen wenigstens ein bestimmter Informationsgewinn erreicht worden ist, während dies für den Mißerfolgstyp nicht der Fall ist. In Tabelle 8 wird gezeigt, wie sich die in beiden Tests empirisch gefundenen Evaluationsergebnisse auf die so definierten Erfolgstypen verteilen.

Tabelle 8: In zwei Delphi-Befragungen ermittelte absolute Häufigkeiten der einzelnen Erfolgstypen bei verschiedenen Schätzaufgaben

Aufgabentyp	Gesamtanzahl der Aufgaben	davon Erfolgstyp 1	2	3	4
4stufige Indikatoren	10	5	4	0	1
2stufige Indikatoren	8	3	0	5	0
Mittelwertschätzungen (7stufig)	36	21	1	11	3
Mittelwertschätzungen (5stufig)	2	1	0	1	0
5stufige Indikatoren	2	0	2	0	0
Gesamt:	58	30	7	17	4

Als Ergebnis zeigt sich, daß insgesamt die folgenden Anteile der einzelnen Typen vorliegen:

- Typ 1 Erfolg: 53 Prozent
- Typ 2 spezifischer Delphi-Erfolg: 12 Prozent
- Typ 3 unspezifischer Delphi-Erfolg: 28 Prozent
- Typ 4 Mißerfolg: 7 Prozent.

Danach hat der Einsatz der Delphi-Methode für die Lösung der gestellten Aufgaben in 93 Prozent der Fälle einen (mehr oder weniger) positiven Verlauf genommen, nur sieben Prozent waren dagegen ausgesprochene Mißerfolge. Die Verteilung der Erfolgstypen auf die einzelnen Aufgabenarten ergibt das folgende in Tabelle 9 gezeigte Bild.

Tabelle 9: In zwei Delphi-Befragungen ermittelte relative Häufigkeiten der einzelnen Erfolgstypen bei verschiedenen Schätzaufgaben

Aufgabentyp	Gesamtanzahl der Aufgaben	davon Erfolgstyp (in Prozent)			
		1	2	3	4
4stufige Indikatoren	10	50	40	0	10
2stufige Indikatoren	8	37,5	0	62,5	0
Mittelwertschätzungen (7stufig)	36	58	3	31	8
Mittelwertschätzungen (5stufig)	2	50	0	50	0
5stufige Indikatoren	2	0	100	0	0
Gesamt	58	53	12	28	7

Es zeigt sich, daß lediglich bei der Schätzung der Antwortverteilungen von vierstufigen Indikatoren und bei der Mittelwertschätzung Mißerfolge (Typ 4) aufgetreten sind.

Direkte Vergleiche dieser Ergebnisse mit anderen aus der Literatur bekannten Evaluationsversuchen sind aufgrund des jeweils unterschiedlichen Designs nicht möglich. Verwiesen werden soll jedoch erstens auf eine Arbeit von Dalkey, der bei zwei Dritteln der gestellten Almanachfragen Verbesserungen der Schätzungen festgestellt hat (vgl. Dalkey 1969a, Dalkey 1969b und Brown et al. 1969). Ein ähnliches Ergebnis hat zweitens das Experiment von Brown und Helmer erbracht. Auch sie stellen in zwei Dritteln der gestellten Fragen Verbesserungen und in einem Drittel Verschlechterungen der Schätzungen fest (vgl. Brown/ Helmer 1964). Diese Ergebnisse stellen eine starke Analogie zu den hier gezeigten dar. Eine Zusammenfassung der Typen 1 und 2 als „Delphi-Erfolg“ würde bedeuten, daß dieser hier ebenfalls in etwa zwei Dritteln der Fälle (bei 65 Prozent) auftritt.

3.3. Delphi oder eine alternative Mehode?

Einer mitunter geäußerten kritischen Auffassung (vgl. z. B. Seeger 1979, S. 150 und Woudenberg 1991, S. 131ff.) entsprechend sei die Delphi-Methode noch solange mit einem beträchtlichen Legitimationsdefizit behaftet, bis sie sich gegenüber anderen Gruppenverfahren als deutlich überlegen abgehoben habe. Vor dem Hintergrund der referierten Ergebnisse soll dieser Auffassung in ihrer Konsequenz widersprochen werden.

Die Ergebnisse der beiden vorgestellten Tests unterstützen eine These, nach der es sich bei der Delphi-Methode um ein gleichberechtigtes sozialwissenschaftliches Instrument zur Gewinnung empirischer Daten handelt. Es kann als nachgewiesen gelten, daß die Delphi-Methode wesentliche Informationen zur Aufklärung ganz bestimmter Fragestellungen zu liefern vermag. Die konkrete Entscheidung für die Benutzung der Delphi-Methode (oder für ein anderes Verfahren) ergibt sich aus dem jeweiligen Forschungszusammenhang. Um hier jedoch eine sachkundige Entscheidung treffen zu können, müssen die Möglichkeiten und Grenzen von Delphi genauer dargestellt werden als bisher.

Kreutz beschreibt noch relativ vage das Einsatzgebiet der Delphi-Technik: „In einem umfassenden Bezugsrahmen gesehen, kann man in der Delphi-Technologie ein Verfahren der Meinungstechnologie (der Ausdruck ‘opinion technology’ wird von N. Dalkey gebraucht) sehen, die sich vor allem dazu eignet, gegenüber offenen Problemsituationen Gruppenleistungen vom Typus des Bestimmens zu initiieren. Sie trägt somit zur ‘gesellschaftlichen Konstruktion der Realität’ bei, indem sie eine Vereinheitlichung der Auffassungen und eine zumindest teilweise Ausschaltung von Widersprüchen erreicht“ (1972, S. 150). Eine Zusammenstellung von Möglichkeiten und Grenzen der Delphi-Methode ermöglicht es dagegen besser abzuschätzen, in welchen Fällen die Anwendung dieses Verfahrens angebracht ist.

Als Argumente *gegen* den Delphi-Ansatz sind zu nennen:

- Es besteht teilweise der Eindruck einer gewissen Zufälligkeit, mit der die Delphi-Methode Ergebnisse gewinnt. Seeger verlangt in diesem Zusammenhang, den Hintergrund offenzulegen, auf dem die Expertengruppe zu ihren Urteilen gelangt (1979, S. 151) sowie die kognitionspsychologischen Grundlagen weiter auszuarbeiten. Weiterhin muß auch eine relativ große Willkür bei der konkreten Gestaltung des Designs einer Delphi-

Erhebung konstatiert werden. Auf professionelle Standards für die Anwendung des Delphi-Verfahrens oder gar auf einen getesteten Regelkanon kann der Anwender dieser Methode (noch) nicht zurückgreifen.

- Vor dem genannten Hintergrund des Mangels an Standards sowie wegen des bei der Anwendung von Delphi üblichen Variantenreichtums fällt es dem Anwender schwer, eine notwendige Fehlerabschätzung vorzunehmen.
- Im Vergleich zu direkten Gruppeninteraktionen ist das (traditionelle) Delphi-Verfahren aufgrund des für die schriftliche Befragung erforderlichen Aufwandes (Rücklauf) zeitintensiv.

Für die Anwendung der Delphi-Methode spricht wiederum:

- Das Problem der Erreichbarkeit (wirklicher) Experten läßt es lohnenswert erscheinen, die Delphi-Technik anzuwenden. So behaupteten auf eine entsprechende Frage hin beispielsweise von den 2300 ursprünglich 1995 im Rahmen eines Prognose-Delphis des Bundesministeriums für Bildung, Wissenschaft, Forschung und Technologie angeschriebenen Personen lediglich zehn von sich, für bestimmte Teilgebiete (vor allem in der Klimaforschung und -technologie) über „große Fachkenntnis“ zu verfügen. Ein ähnliches Ergebnis wurde in Japan erzielt (vgl. Cuhls et al. 1995). Es erscheint zweifelhaft, daß es gelänge, eine solche exklusive Expertengruppe anders als über eine Delphi-Befragung zu rekrutieren, zumindest wenn der dabei erforderliche Aufwand mit in Rechnung gestellt wird.
- Die Delphi-Methode bietet gerade dort eine Lösung, wo viele andere Methoden versagen. So kann das Delphi-Verfahren gewählt werden, um die Ergebnisse von Bevölkerungsbefragungen zu schätzen; beispielsweise dann, wenn die Zielpopulation nicht für Befragungen zu erreichen ist, es sich um retrospektive Sachverhalte handelt oder besonders komplexe Probleme interessieren.

Die in diesem Aufsatz referierten Ergebnisse zweier Tests zur Schätzung von Befragungsergebnissen sollten dazu dienen, die Leistungsfähigkeit des Delphi-Ansatzes für genau diese sehr spezielle Fragestellung zu demonstrieren. Ausdrücklich nicht Gegenstand der Betrachtungen war es, verschiedene Methoden vergleichend gegenüberzustellen. Obwohl zwar gezeigt werden konnte, daß Delphi zur Aufklärung bestimmter Fragestellungen beizutragen vermag, halten die auf diese Weise gefundenen Ergebnisse einem Vergleich mit den „tatsächlichen“ Daten nicht stand. Wenn jedoch keine Möglich-

keit für eine direkte Befragung der Zielpopulation besteht, stellt die Delphi-Technik einen methodisch vertretbaren Ausweg bei der Beschaffung notwendiger Informationen dar. Dies gilt insbesondere dann, wenn ihre Anwendungsbedingungen weiter ausgearbeitet werden. Es wäre damit auch völlig verfehlt, etwa die Frage zu stellen, ob nicht Delphi-Erhebungen einmal Bevölkerungsbefragungen ersetzen können. Eine solches Anliegen steht in keiner Beziehung zu der hier verfolgten Zielstellung: die Methodik von Delphi-Erhebungen weiterzuentwicklen.

4. Zusammenfassung

In der Literatur werden sehr unterschiedliche Kriterien für die Beurteilung einer Delphi-Befragung benutzt. In diesem Aufsatz wurden ebenfalls verschiedene Gesichtspunkte aufgegriffen bzw. entwickelt und für die Bewertung eines spezifischen Delphi-Ansatzes - die Schätzung von Ergebnissen einer Bevölkerungsbefragung - benutzt. Das Ergebnis erbrachte zunächst unterschiedliche, mitunter widersprüchliche Aussagen. Kein Kriterium und damit kein einzelnes Ergebnis konnte *die* Hauptaussage zum Funktionieren von Delphi liefern. Alle Kriterien enthalten indes spezifische Anhaltspunkte zur Evaluation von Delphi-Befragungen. Der Versuch, verschiedene Gesichtspunkte zu kombinieren, erbrachte synthetisch gebildete „Erfolgstypen". Die Tragfähigkeit dieses Ansatzes ist in weiteren Tests zu überprüfen.

Insgesamt ist festzustellen, daß die Ergebnisse der beiden Tests sich in eine ganze Reihe an vorliegenden Untersuchungsergebnissen einordnen lassen, die belegen, daß prinzipiell mit Hilfe von Delphi-Befragungen ein gewisses Maß an Unklarheit beseitigt und Informationsgewinn erzielt werden kann.

Versuche, die Delphi-Methode anderen Verfahren als Alternative gegenüberzustellen, sind als nicht besonders produktiv erkannt worden. Sinnvoller erscheint es dagegen, nach Möglichkeiten zu suchen, um die Delphi-Technik so zu verfeinern, daß sie für den jeweiligen konkreten Anwendungszweck zu optimalen Ergebnissen führt. Da in einem betrachteten Beispiel immerhin (vgl. Abschnitt 3.2.1.) etwa ein Drittel der Experten über besseres Wissen verfügt als der Gruppendurchschnitt, kann davon ausgegangen werden, daß es noch einen gewissen Spielraum für die Weiterentwicklung des Delphi-Designs gibt.

Das in den beiden Tests für die Beurteilung von Delphi gewählte konkrete Anwendungsgebiet - die Schätzung von Ergebnissen von Bevölkerungsbefragungen - erscheint durchaus von Bedeutung für die Sozialwissenschaft. Im Rahmen der AIDS-Forschung wurden beispielsweise Prostituierte über deren Freier befragt. Da kein direkter empirischer Zugang zur letzteren Gruppe gefunden werden konnte, mußte ein solcher Weg gewählt werden (vgl. Markert 1994). Auch heikle Fragestellungen, besonders

komplizierte Probleme oder die retrospektive Gewinnung von Informationen wären sozialwissenschaftlich wichtige Anwendungsmöglichkeiten eines solchen Delphi-Ansatzes.

In weiteren Untersuchungen erscheint es vor allem notwendig, die folgenden Themen zu bearbeiten:

- Systematische Studien, die die kognitionspsychologischen Grundlagen der Erkenntnisgewinnung bei einer Delphi-Befragung weiter aufdecken.
- Die Entwicklung von Gewichtungsstrategien für die Expertenurteile. Wenn es gelingt nachzuweisen, z.B. mit Hilfe von in einer Delphi-Befragung aufzunehmender Kontrollindikatoren, inwieweit die Experten dazu in der Lage sind, ihre eigene Kompetenz richtig zu beurteilen, würde es möglich werden, daraus entsprechende Gewichtungsfaktoren abzuleiten.
- Die Ausarbeitung von Standards, nach denen bei der Erstellung eines konkreten Erhebungsdesigns für Delphi vorzugehen ist. Solche Standards sollten vor allem für die Rekrutierung der geeigneten Experten und die konkrete Gestaltung der Erhebung ausgearbeitet werden.
- Eine möglichst exakte Eingrenzung jener Problembereiche, für die ein Einsatz der Delphi-Technik angemessen ist.

Literatur

Albach, H., 1970: Informationsgewinnung durch strukturierte Gruppenbefragung. Die Delphi-Methode. In: Zeitschrift für Betriebswirtschaft 40/1970 (Ergänzungsheft): 11-26.

Becker, D., 1974: Analyse der Delphi-Methode und Ansätze zu ihrer optimalen Gestaltung. Inaugural - Dissertation zur Erlangung der Würde eines Doktors der Wirtschaftswissenschaften der Universität Mannheim.

Brockhoff, K., 1979: Delphi-Prognosen im Computerdialog. Experimentelle Erprobung und Auswertung kurzfristiger Prognosen, Tübingen: J. C. Mohr.

Brown, B./ Helmer, O., 1964: Improving the Reliability of Estimates Obtained from a Consensus of Experts, RAND Corporation, P-2986.

Brown, B./ Cochran, S./ Dalkey, N., 1969: The Delphi Method II. Structure of Experiments, RAND Corporation, RM-5957-PR.

BMFT, 1993: Deutscher Delphi-Bericht zur Entwicklung von Wissenschaft und Technik, im Auftrag des Bundesministeriums für Forschung und Technologie (BMFT), Bonn.

Cuhls, K./ Breiner, S./ Grupp, H., 1995: Delphi-Bericht 1995 zur Entwicklung von Wissenschaft und Technik. Endbericht an das Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF), Frauenhofer-Institut für Systemtechnik und Innovationsforschung, Karlsruhe.

Dalkey, N. C., 1969a: The Delphi Method: An Experimental Study of Group Opinion. RAND RM 5888-PR, June.

Dalky, N. C., 1969b: Analyses from a Group Opinion Study, in: Futures I, S. 541-551.

Erffmeyer, R. C./ Lane, I. M., 1984: Quality and Acceptance of an Evaluative Task: The Effects of Four Group Decision-Making Formats, in: Group and Organization Studies 9, S. 509.529.

Fischer, G. W., 1981: When Oracles Fail: A Comparision of Four Procedures for Aggregating Subjective Probability Forecasts, in: Organizational Behavior and Human Performance 28, S. 96 - 110.

Geschka, H., 1977: Delphi. In: Bruckmann, G. (Hrsg.), Langfristige Prognosen. Möglichkeiten und Methoden der Langfristprognostik komplexer Systeme. Würzburg/Wien.

Häder, M./ Häder, S., 1994a: Die Grundlagen derDelphi-Methode. Ein Literaturbericht, ZUMA-Arbeitsbericht Nr. 94/02.

Häder, M./ Häder, S., 1994b: Ergebnisse einer experimentellen Studie zur Delphi-Methode, Mannheim, ZUMA-Arbeitsbericht Nr. 94/05.

Häder, M./ Häder, S./ Ziegler, A., 1995: Punkt- vs. Verteilungsschätzungen: Ergebnisse eines Tests zur Validierung der Delphi-Methode, Mannheim, ZUMA-Arbeitsbericht Nr. 95/05.

Häder, M./ Häder,S., 1995: Delphi und Kognitionspsychologie: Ein Zugang zur theoretischen Fundierung der Delphi-Methode, in: ZUMA-Nachrichten 37. S. 8-34.

Hell, W./ Fiedler, K./ Gigerenzer, G. (Hrsg.), 1993: Kognitive Täuschungen. Fehl-Leistungen und Mechanismen des Urteilens, Denkens und Erinnerns, Spektrum Akademischer Verlag. Heidelberg Berlin Oxford.

Hertwig, R., 1993: Frequency-Validity-Effekt und Hindsight-Bias: Unterschiedliche Phänomene - gleiche Prozesse? In: Hell et al. 1993. S. 39-71.

Hoffrage, U., 1993: Die Illusion der Sicherheit bei Entscheidungen unter Unsicherheit. In: Hell et al. 1993. S. 73-97.

Jugend ’92, Maschinenlesbares Codebuch ZA Nr. 2323, Zentralarchiv für Empirische Sozialforschung an der Universität zu Köln.

Köhler, G., 1992: Methodik und Problematik einer mehrstufigen Expertenbefragung, in: J. H. P. Hoffmeyer-Zlotnik (Hrsg.), Analyse verbaler Daten. Über den Umgang mit qualitativen Daten. Opladen: Westdeutscher Verlag.

Kreutz, H., 1972: Soziologie der empirischen Sozialforschung. Stuttgart: Enke.

Linestone, H.A./ Turoff, M. (Hrsg.), 1975: The Delphi Method. London u.a.: Addison-Wesley.

Markert, St., 1994: Risikoverhalten von Freiern, in: Heckmann, W./ Koch, M. A. (Hrsg.), Sexualverhalten in Zeiten von AIDS. Berlin. S. 369 - 375.

Seeger, Th., 1979: Die Delphi-Methode. Expertenbefragungen zwischen Prognose und Gruppenmeinungsbildungsprozessen; überprüft am Beispiel von Delphi-Befragungen im Gegenstandsbereich Information und Dokumentation. Diss., Freiberg: Hochschul-Verlag.

Woudenberg, F., 1991: An Evaluation of Delphi, in: Technological Forecast and Social Change 40, S. 131 - 150.

Anhang: Zusammenstellung der in den beiden Tests benutzten Indikatoren[11]

1. Frage

Die meisten Menschen gehen ja ganz unterschiedlich mit der Zeit in ihrem Leben um und planen auch unterschiedlich. Wie gut beschreiben die Sätze auf diesen Kärtchen Deine Meinung?

Heute ist heute und morgen ist morgen.

- überhaupt nicht
- weniger gut
- gut
- sehr gut

2. Frage (gleicher Fragetext)

Was ich nächste Woche machen werde, überlege ich mir dann, wenn es soweit ist.

- überhaupt nicht
- weniger gut
- gut
- sehr gut

3. Frage (gleicher Fragetext)

Ich tue am liebsten spontan das, wozu ich gerade Lust habe.

- überhaupt nicht
- weniger gut
- gut
- sehr gut

4. Frage (gleicher Fragetext)

Ich höre gerne Geschichten aus guten alten Zeiten.

- überhaupt nicht

[11] Alle Indikatoren wurden der Shell-Jugendstudie ‘92 entnommen (vgl. Jugend ‘91)

- weniger gut
- gut
- sehr gut

5. Frage (gleicher Fratetext)

Im Leben ist alles Zufall.

- überhaupt nicht
- weniger gut
- gut
- sehr gut

6. Frage

Wie stehst Du zur Vereinigung von ehemaliger DDR und alter Bundesrepublik von heute aus gesehen?

1 sehr dafür

2 eher dafür als dagegen

3 unentschieden

4 eher dagegen als dafür

5 sehr dagegen

7. Frage

Die Menschen sind ja sehr unterschiedlich, wenn es um Ihre Lebensziele geht. Manche sind sehr anspruchsvoll und ehrgeizig, andere finden diese weniger gut oder wichtig. Wie ist das bei Dir?

Ich bin bei den Zielen, die ich mir für mein Leben setze ...

1 nicht so anspruchsvoll oder ehrgeizig

2

3

4

5 sehr anspruchsvoll und ehrgeizig

8. Frage

Auf diesen Karten stehen jeweils unterschiedliche Meinungen, wie man mit seinem Leben umgehen kann. Sage mir bitte, welcher Meinung Du eher zustimmst.

A1 Ich betrachte mein Leben als eine Aufgabe, für die ich da bin und für die ich alle Kräfte einsetze. Ich möchte in meinem Leben etwas leisten, auch wenn das oft schwer und mühsam ist.

A2 Ich möchte mein Leben genießen und mich nicht mehr abmühen als nötig. Man lebt schließlich nur einmal, die Hauptsache ist doch, daß man etwas von seinem Leben hat.

9. Frage (gleicher Fragetext)

B1 Ich finde es wichtig, mein Leben so einzurichten, daß ich ein ganz anderer Mensch bin, der anders ist als alle anderen Menschen in meiner Umgebung.

B2 Ich finde es wichtig, mein Leben so einzurichten, daß betont wird, was ich mit anderen Menschen um mich herum gemeinsam habe und worin wir uns ähnlich sind.

10. Frage (gleicher Fragetext)

C1 Ich finde es am wichtigsten, daß ich im Leben selbständig bin und wirklich selbständig meine eigenen Interessen und Ziele verfolge.

C2 Ich finde es am wichtigsten, daß ich im Leben und meinen Entscheidungen die Interessen und Ziele anderer Menschen berücksichtige.

11. Frage (gleicher Fragetext)

D1 Ich richte mein Leben so ein, daß die Dinge, die ich tue, mir sogleich und direkt etwas bringen und daß ich unmittelbar einen Nutzen sehe oder Spaß daran habe.

D2 Ich richte mein Leben so ein, daß die Dinge, die ich tue, sich langfristig auszahlen, daß ich später im Leben einmal etwas davon habe und auf lange Frist die Früchte meines Tuns ernten kann.

12. Frage

Auf dieser Liste stehen einige Dinge, die wichtig als Werte für das eigene Leben sein können, was man anstrebt und wie man leben möchte. Sage Du mir bitte bei jedem, wie wichtig es Dir ist für Dein Leben.

1 = wäre nicht wichtig und 7 = wäre äußerst wichtig

01. Innere Harmonie (in Frieden mit mir selbst)
02. Soziale Macht (Kontrolle über andere, Dominanz)
03. Freiheit (Freiheit des Handelns und des Denkens)
04. Soziale Ordnung (Stabilität der Gesellschaft)
05. Ein anregendes Leben (anregende Erfahrungen)
06. Höflichkeit (gute Umgangsformen)
07. Reichtum (materieller Besitz, Geld)
08. Nationale Sicherheit (Schutz meiner Nation gegen Feinde)
09. Kreativität (Originalität, Phantasie)
10. Eine Welt in Frieden (frei von Krieg und Konflikt)
11. Achtung vor Tradition (Erhaltung ehrwürdiger Sitten)
12. Loslösung (von weltlichen Belangen)
13. Familiäre Sicherheit (Sicherheit für die geliebten Personen)
14. Einheit mit der Natur (Einpassung in die Natur)
15. Ein abwechslungsreiches Leben (erfüllt mit Herausforderungen, Neuem und Veränderungen)
16. Autorität (ein Recht zu führen und zu bestimmen)
17. Wahre Freundschaft (enge unterstützende Freunde)
18. Eine Welt der Schönheit (Schönheit der Natur und Künste)